Predicting Transcription Factor Binding Sites Using Structural Knowledge

Tommy Kaplan[1,2], Nir Friedman[1,⋆], and Hanah Margalit[2,⋆]

[1] School of Computer Science,
The Hebrew University, Jerusalem 91904, Israel
{tommy, nir}@cs.huji.ac.il
[2] Dept. of Molecular Genetics and Biotechnology,
Hadassah Medical School, The Hebrew University,
Jerusalem 91120, Israel
hanah@md.huji.ac.il

Abstract. Current approaches for identification and detection of transcription factor binding sites rely on an extensive set of known target genes. Here we describe a novel structure-based approach applicable to transcription factors with no prior binding data. Our approach combines sequence data and structural information to infer context-specific amino acid-nucleotide recognition preferences. These are used to predict binding sites for novel transcription factors from the same structural family. We apply our approach to the Cys$_2$His$_2$ Zinc Finger protein family, and show that the learned DNA-recognition preferences are compatible with various experimental results. To demonstrate the potential of our algorithm, we use the learned preferences to predict binding site models for novel proteins from the same family. These models are then used in genomic scans to find putative binding sites of the novel proteins.

1 Introduction

Specific binding of transcription factors to cis-regulatory elements is a crucial component of transcriptional regulation. Previous studies have used both experimental and computational approaches to determine the relationships between transcription factors and their targets. In particular, probabilistic models were employed to characterize the binding preferences of transcription factors, and to identify their putative sites in genomic sequences [24, 27]. This approach is useful when massive binding data are available, but it cannot be applied to proteins without extensive experimental binding studies. This difficulty is particularly emphasized in view of the genome projects, where new proteins are classified as DNA-binding according to their sequence, yet there is no information about the genes they regulate.

To address the challenge of profiling the binding sites of novel proteins, we propose a family-wise approach that builds on structural information and on

[⋆] Correspondence authors.

S. Miyano et al. (Eds.): RECOMB 2005, LNBI 3500, pp. 522–537, 2005.

the known binding sites of other proteins from the same family. We use solved protein-DNA complexes [16, 18] to determine the exact architecture of interactions between nucleotides and amino acids at the DNA-binding domain. Although sharing the same structure, different proteins from a structural family obtain different binding specificities due to the presence of different residues at the DNA-binding positions. To predict their binding site motifs, we need to identify the residues at these key positions and understand their DNA-binding preferences.

In previous studies, we used the empirical frequencies of amino acid-nucleotide interactions [17–19] in solved protein-DNA complexes (from various structural families) to build a set of *DNA-recognition preferences*. This approach assumed that an amino acid has common nucleotide-binding preferences for all structural domains and at all binding positions. However, there are clear experimental indications that this assumption is not always valid: a particular amino acid may have different binding preferences depending on its positional context [9, 10, 14]. To estimate these context-specific DNA-recognition preferences, we need to determine the appropriate context of each residue, which may depend on its relative position and orientation with respect to the nucleotide. For this, we need to collect statistics about the DNA-binding preferences at this context. Naively, this can be achieved from a large ensemble of solved protein-DNA complexes from the same family. Unfortunately, sufficient data of this type are currently unavailable.

To overcome this obstacle, we propose to estimate context-specific DNA-recognition preferences from available sequence data using statistical estimation procedures. The input of our learning algorithm are pairs of transcription factors and their target DNA sequences [27]. We then recognize the specific residues and nucleotides that participate in protein-DNA interaction, and collect statistics about the DNA-binding preferences of residues at different contexts of the binding domain. These preferences can then be used to predict binding sites of other transcription factors from the same family, for which no known targets are available.

2 The Canonical Cys_2His_2 Zinc Finger DNA-Binding Family

We apply our approach to the Cys_2His_2 Zinc Finger DNA-binding family. This family is the largest known DNA-binding family in multi-cellular organisms [26] and has been studied extensively [28]. Many members of this family bind DNA targets according to a stringent binding model [13, 20], which maps the exact interactions between specific residues in the DNA-binding domain along with nucleotides at the DNA site (Figure 1). We term this the *canonical binding model.* In addition, Zinc Finger proteins whose DNA-binding domains are similar to those that bind through this model are termed canonical. According to the canonical binding model the residues involved in DNA binding are located at positions 6, 3, 2, and -1 relatively to the beginning of the α-helix in each fin-

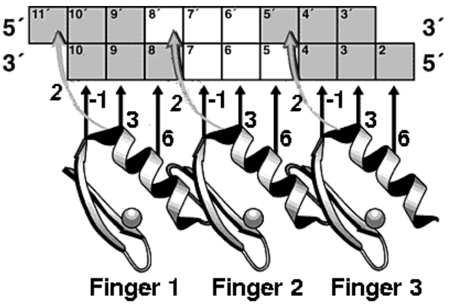


Fig. 1. The canonical Cys_2His_2 Zinc Finger DNA-binding model, based on solved protein-DNA complexes [13, 20]. A protein with three fingers is shown. In each finger, residues at positions 6, 3, 2, and -1 (relatively to the beginning of the α-helix) interact with adjacent nucleotides in the DNA molecule. (Figure adapted from Prof. Aaron Klug, with permission)

ger (Figure 1). Our goal is to extract position-specific amino acid-base binding preferences for each of these positions.

2.1 Sequences of Zinc Finger Proteins and Their DNA Binding Sites

To estimate the recognition preferences, we use the sequences of many Zinc Finger proteins together with their native DNA targets (extracted from the TRANSFAC database [27]). To identify the canonical Cys_2His_2 Zinc Fingers based on their sequence, we trained a profile HMM [12] on 31 experimentally determined canonical domains [28], and used it to classify the remaining Cys_2His_2 Zinc Finger domains in TRANSFAC [27]. From the canonical ones, we only selected proteins with two to four properly spaced fingers. This resulted in 61 canonical proteins, and 455 protein-DNA pairs. We use these as our training data in subsequent steps. The total number of fingers in this dataset was 1320, and the total length of all binding sites was 9761bp (average length of 21bp per site).

2.2 Identification of DNA-Binding Residues

Next, we conceptually "thread" each protein-DNA pair onto the canonical binding model, to obtain an ensemble of residue-nucleotide interactions, from which we can estimate the recognition preferences. To do so, we must first identify the DNA-binding residues. We identify these positions using their relative positioning in the Cys_2His_2 conserved pattern: CX(2-4)CX(11-13)HX(3-5)H. Although theoretically there can be 20^4 different combinations of amino-acids at the four interacting positions, we found only 80 different combinations among the 1320 fingers in our database.

2.3 Identification of DNA Binding Sites

Now that we can identify the interacting residues, we face the problem of identifying the stretch of nucleotides they interact with. Unfortunately, the exact binding locations of the transcription factors are not pinpointed in TRANSFAC, and thus we must employ statistical tools to infer them. In short, we wish to enumerate over all possible alignments of the DNA, and consider the likelihood of each DNA site given the interacting residues. For this we need to use the position-specific amino acid-base recognition preferences that we aim to estimate. We demonstrate in Section 3 how this is achieved, but first let us describe the probabilistic model of the DNA binding site, given that such preferences are available.

2.4 Probabilistic Model for Protein-DNA Interactions

We now consider how to model the DNA binding preferences of a protein given its amino acid sequence. In a probabilistic framework, we describe a model that assigns probabilities for any sequence of nucleotides at the binding site, given the residues they interact with. For a canonical Zinc Finger protein, we denote by $A = \{A_{i,p} : i = \{1, \ldots, k\}, p \in \{-1, 2, 3, 6\}\}$ the set of interacting residues in the different four positions of the k fingers (ordered from the N- to the C-terminus). Let $N_1, \ldots, N_L$ be a target DNA sequence. The conditional probability of an interaction with a DNA subsequence, starting from the j'th position in the DNA, is:

$$P(N_j, \ldots, N_{j+3k-1}|A) = \prod_{i=1}^{k} P_6(N_{j+3(i-1)}|A_{k+1-i,6}) P_3(N_{j+3(i-1)+1}|A_{k+1-i,3}) \\ P_{-1}(N_{j+3(i-1)+2}|A_{k+1-i,-1}) \qquad (1)$$

where $P_p(N|A)$ is the conditional probability of nucleotide N given amino acid A at position p. These probabilities are the parameters of the model. For each of the four interacting positions, there should be a matrix of the conditional probabilities of the four nucleotides given all 20 residues. We call these matrices the *DNA-recognition preferences.*

We model the non-interacting nucleotides in the sequence using a background model P_{BG} (e.g. a uniform mononucleotide model), thus the probability of a sequence of length L that contains a binding site at position j is:

$$P(N|A, j) = P_{BG}(N_1, \ldots, N_{j-1}) P(N_j, \ldots, N_{j+3k-1}|A) P_{BG}(N_{j+3k}, \ldots, N_L) \qquad (2)$$

Since the exact positioning j of the binding site is not known, we enumerate over all possible values:

$$P(N|A) = \sum_j P(j) P(N|A, j) \qquad (3)$$

where $P(j)$ is the prior probability of binding at position j. To handle truncated sites in the TRANSFAC database, we allow j to range between $-3/4 * 3k$ (when only the last quarter of the binding site is present) and $L - 3/4 * 3k + 1$ (only first quarter is present). We use a uniform prior over the j's, with the exception of missing nucleotides, which are penalized exponentially.

The model, as described above, does not account for interactions by the amino acids in positions 2 in each finger nor reverse complement binding. The latter requires only minimal adjustments, and is handled using an additional orientation variable. Handling the residues at position 2 is a bit trickier. According to the canonical binding model (Figure 1), the amino acid at position 2 interacts with the nucleotide that is complementary to the nucleotide interacting with position 6 of the previous finger. Thus, when we have a base pair interacting with two amino acids, we replace the term $P_6(N_{j+3(i-1)}|A_{k+1-i,6})$ by a term:

$$\alpha P_6(N_{j+3(i-1)}|A_{k+1-i,6}) + (1-\alpha)P_2(N_{j+3(i-1)}|A_{k+2-i,2}) \qquad (4)$$

for $i > 1$, where α is a weighting coefficient that depends on the number of samples seen while estimating the recognition preferences at each position. Moreover, we add the term $P_2(N_{j+3(i-1)}|A_{k+2-i,2})$, for $i = k+1$, to capture the last nucleotide, that is in interaction with position 2 of the first finger.

3 Learning DNA-Recognition Preferences from Sequence Data

We use the sequences of the proteins and their target DNA sites, and estimate four sets of position-specific DNA-recognition preferences that maximize the likelihood of the DNA given the binding proteins. As stated above, although the DNA sequences in our database were reported as bound by their corresponding proteins [27], the exact binding locations are not documented. Thus, we need to simultaneously identify the exact binding locations and optimize the parameters of DNA-recognition. For this, we use the iterative *Expectation Maximization* (EM) algorithm [11]. We start with some initial choice of DNA-recognition preferences (possible choices are discussed below). The algorithm proceeds iteratively, by carrying out these two steps, as illustrated in Figure 2.

E-step: For every protein-DNA pair, we compute the expected posterior probability that the binding begins in position j, using the current sets of DNA-recognition preferences θ_t.

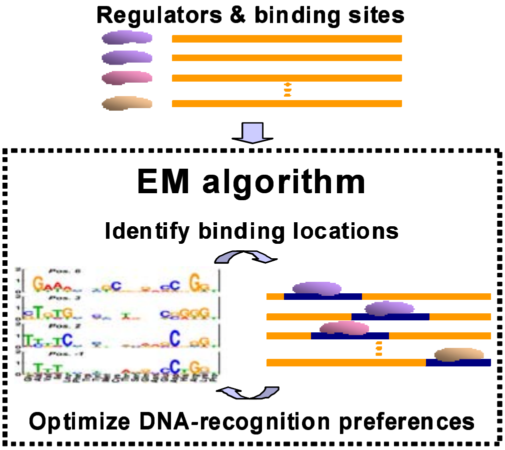

Regulators & binding sites
EM algorithm
Identify binding locations
Pos. 6
Pos. 3
Pos. 2
Pos. -1
Optimize DNA-recognition preferences

Fig. 2. Estimating the DNA-recognition preferences. The recognition preferences are estimated from unaligned pairs of transcription factors and their DNA targets from the TRANSFAC database [27] (shown on top). The EM algorithm is used to simultaneously assess the exact binding positions of each protein-DNA pair (bottom-right), and to estimate four sets of position-specific DNA-recognition preferences (bottom-left)

$$P(j|A,N) = \frac{P_{\theta_t}(N|A,j)}{\sum_{j'} P_{\theta_t}(N|A,j')} \tag{5}$$

M-step: Next, we update the sets of DNA-recognition preferences θ_{t+1} to maximize the likelihood of the current binding positions j's for all protein-DNA pairs. This is based on the posterior probabilities that were computed in the E-step. Specifically, the conditional probability $P_p(n|a)$ of each nucleotide n given amino acid a at position p of the Zinc finger domain is estimated in θ_{t+1} using the expected number of interactions between n and a at p over all protein pairs, given the posterior probabilities j's.

The EM algorithm is proved to converge, since each of these two steps increases the likelihood of the data [11]. Obviously, this does not ensure that the final sets of DNA-recognition preferences θ_T are the *optimal* ones, due to sub-optimal local maxima of the likelihood function. This can be overcome by applying the EM procedure with multiple random starting points or by using prior knowledge starting points. An additional potential pitfall is over-fitting the recognition preferences of rare residues. To address this problem and ensure that the estimated parameters for rare amino-acids are close to uniform distribution (i.e., uninformative), we use a standard method of regularization by *pseudo-counts*. By applying a uniform *Dirichlet* prior, we add a constant (0.7 in the results below) to each amino acid-nucleotide count computed at the end of the E-step. We then perform a maximum *a-posteriori* estimation rather than maximum likelihood estimation.

We evaluate the robustness and convergence rate of the EM algorithm using a 10-fold cross validation procedure. In each round, we remove part of the data, train on the remaining pairs, and test the likelihood of the held-out protein-DNA pairs. We use this procedure to test various initialization options, including random starting points, and the general protein-DNA recognition preferences that were learned from all protein-DNA families [17]. Figure 3 shows the average likelihood per interaction on held-out test data, for various starting points. As shown, the EM algorithm performs best when initialized with the general recognition preferences, converging within few iterations. Similar likelihood results were ob-

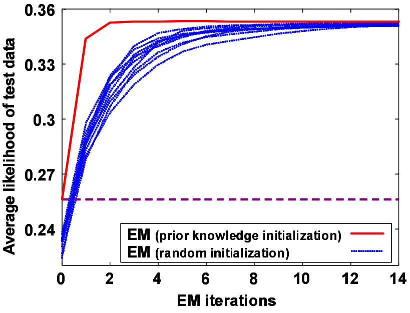


Fig. 3. 10-fold cross-validation tests show the average likelihood per interaction (in bits) of held-out test data. We show the likelihood along 14 EM iterations using various starting points. The thick red line marks the likelihood obtained when starting from the general set of DNA-recognition preferences [18]. Random starting points are plotted using thin blue lines. The likelihood of the data according to the general set of DNA-recognition preferences [17] is shown with the horizontal dashed line

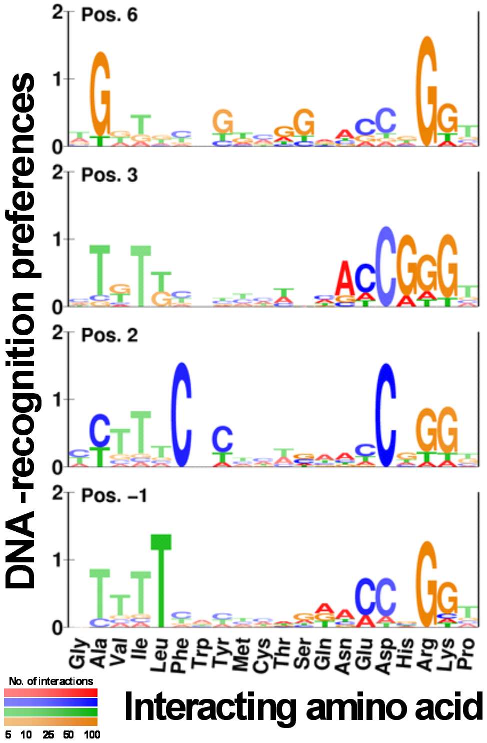


Fig. 4. Four sets of position-specific DNA-recognition preferences for canonical Cys_2His_2 Zinc Fingers. The estimated sets of DNA-recognition preferences for the DNA-binding residues at positions 6, 3, 2, and -1 of the Zinc Finger domain are displayed as sequence logos. At each position, the associated distribution of nucleotides is displayed for each amino acid. The total height of letters represents the information content (in bits) of the position, and the relative height of each letter represents its probability. Color intensity indicates the level of confidence for the preferences of a given amino acid at a certain position (where pale colors indicate low confidence positions due to a small number of occurrences of the residue at the specific position within the training data). Some of the DNA-binding preferences are general, regardless of the residue's position within the Zinc Finger domain (e.g. the tendency of lysine to bind guanine (G)), while others are position-dependent (e.g. the tendency of phenylalanine to bind cytosine only when in position 2)

tained using random starting points, although the convergence rate is somewhat slower. Figure 3 also shows that the algorithm does not over-fit the training data, as this would have led to deteriorated performances over the held-out test data. The optimized sets of position-specific DNA-recognition preferences (estimated from full training data) are shown in Figure 4.

3.1 Recognition Preferences Are Consistent with Experimental Results

We evaluated the four sets of DNA-recognition preferences by comparing them with experimental data. First, we compared the derived preferences with qualitative preferences based on phage-display experiments [28] and found the two to be consistent. Second, we predicted binding site models for various variants of the Egr-1 protein, for which experimental binding data were available using DNA-microarrays [6]. We then used the predicted binding models to score each

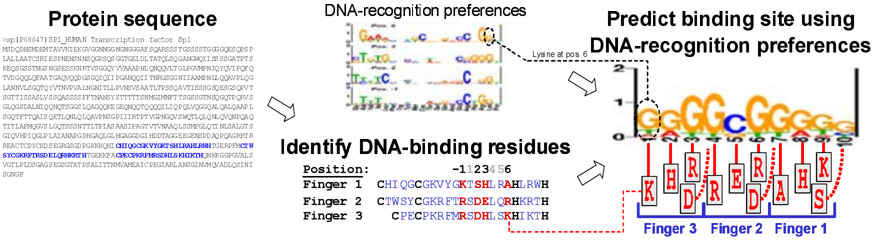
Table 1. Correlation with experimentally measured binding affinities. We compare the ranking of the predicted binding sites to the experimental binding results of Bulyk et al. [6, 7], using Spearman rank correlation. Oligonucleotides with low binding affinities (baseline noise value) were considered as non-viable and not taken into account

Variant of Egr-1	Spearman correlation coefficient	number of viable oligos	p-value
wt	0.73	15	<0.0025
LRHN	0.60	12	<0.025
REDV	0.83	15	<0.0005
RGPD	0.67	17	<0.0025

of the possible DNA binding sites that were tested in the experimental study. We found that our predictions were significantly correlated with the experimentally measured binding affinities (Table 1).

3.2 Predicting Binding Sites of Novel Proteins Within Genomic Sequences

We now turn to evaluate the ability of the estimated DNA-recognition preferences to identify the binding sites of novel proteins within genomic sequences.


Protein sequence
DNA-recognition preferences
Lysine at pos. 6
Predict binding site using DNA-recognition preferences
2
1
0
Identify DNA-binding residues
K
H
R
D
R
E
R
D
A
H
K
S
Finger 3
Finger 2
Finger 1

```
>sp|P08047|SP1_HUMAN Transcription factor Sp1
MSDQDHSMDEMTAVVKIEKGVGGNNGGNGNGGGAFSQARSSSTGSSSSTGGGGQESQPSP
LALLAATCSRIESPNENSNNSQGPSQSGGTGELDLTATQLSQGANGWQIISSSSGATPTS
KEQSGSSTNGSNGSESSKNRTVSGGQYVVAAAPNLQNQQVLTGLPGVMPNIQYQVIPQFQ
TVDGQQLQFAATGAQVQQDGSGQIQIIPGANQQIITNRGSGGNIIAAMPNLLQQAVPLQG
LANNVLSGQTQYVTNVPVALNGNITLLPVNSVSAATLTPSSQAVTISSSGSQESGSQPVT
SGTTISSASLVSSQASSSSFFTNANSYSTTTTTSNMGIMNFTTSGSSGTNSQGQTPQRVS
GLQGSDALNIQQNQTSGGSLQAGQQKEGEQNQQTQQQQILIQPQLVQGGQALQALQAAPL
SGQTFTTQAISQETLQNLQLQAVPNSGPIIIRTPTVGPNGQVSWQTLQLQNLQVQNPQAQ
TITLAPMQGVSLGQTSSSNTTLTPIASAASIPAGTVTVNAAQLSSMPGLQTINLSALGTS
GIQVHPIQGLPLAIANAPGDHGAQLGLHGAGGDGIHDDTAGGEEGENSPDAQPQAGRRTR
REACTCPYCKDSEGRGSGDPGKKKQHICHIQGCGKVYGKTSHLRAHLRWHTGERPFMCTW
SYCGKRFTRSDELQRHKRTHTGEKKFACPECPKRFMRSDHLSKHIKTHQNKKGGPGVALS
VGTLPLDSGAGSEGSGTATPSALITTNMVAMEAICPEGIARLANSGINVMQVADLQSINI
SGNGF
```

```
Position:            -1123456
Finger 1  CHIQGCGKVYGKTSHLRAHLRWH
Finger 2  CTWSYCGKRFTRSDELQRHKRTH
Finger 3    CPECPKRFMRSDHLSKHIKTH
```

Fig. 5. Predicting the DNA binding site motifs of novel transcription factors. Given the sequence of a novel protein (shown on left), its DNA-binding domains (in blue) are identified using the Cys_2His_2 conserved pattern. The residues at the key positions (6, 3, 2 and -1) of each finger (marked in red in the middle-bottom panel) are then assigned onto the canonical binding model (on right), and the sets of position-specific DNA-recognition preferences (middle-top panel) are used to construct a probabilistic model of the DNA binding site (right). For example, position 1 in the binding site is determined by the binding preferences of the lysine (K) at the sixth position of the third finger (dotted red and black lines). We predict the nucleotide probabilities at this position using the appropriate recognition preferences (dotted black line). A web-server for predicting the binding sites of Cys_2His_2 Zinc Finger proteins can be accessed at `http://compbio.cs.huji.ac.il/predictBS`

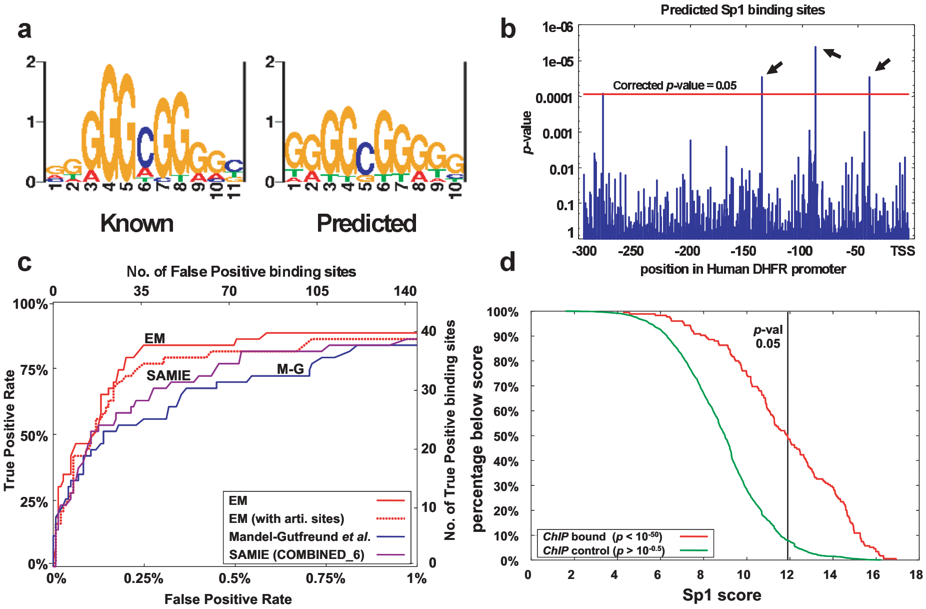


Fig. 6. Validation of the DNA-recognition preferences. (a) The predicted binding site model of human Sp1 protein is compared to its known site (matrix V$SP1_Q6 from TRANSFAC [27], based on 108 aligned binding sites). To prevent bias by known Sp1 sites in our training data, we applied a "leave-protein-out" cross-validation approach, and predicted the DNA-binding model of Sp1 using DNA-recognition preferences that were learned from a subset of the training data, after removing all Sp1 sites. (b) We then scanned the 300bp-long promoter of human dihydrofolate reductase (DHFR) using the predicted Sp1 binding site model using the CIS program [2]. The p-value of each potential binding site is shown (y-axis). Four positions achieved a significant p-value (≤ 0.05) after a Bonferroni correction for multiple hypotheses (red horizontal line). Out of these four, three are known Sp1 binding sites [15] (marked by arrows). (c) A summary of the *in silico* binding experiments for natural 43 known binding sites. Shown is the tradeoff between False Positive rate (x-axis) and True Positive rate (y-axis) as the threshold for putative binding sites is changed, using an ROC curve. For every threshold point, our sets of recognition preferences (marked EM) achieve higher accuracy than the preferences of Mandel-Gutfreund et al. [17] (marked M-G) and Benos et al. [4] (marked SAMIE). Interestingly, when the DNA-recognition preferences were estimated from training data that were expanded to include artificial sequences, from TRANSFAC we obtained inferior results (dotted red line). (d) Comparison of cumulative distributions of Sp1 logodd scores within genomic sequences of Sp1 targets and non-targets determined by unbiased chromatin immunoprecipitation (*ChIP*) scans of human Chromosomes 21 and 22 [8]. The predicted Sp1 motif appears in a significant manner ($p \leq 0.05$) in 45% of the target sequences but only in 5% of the control sequences

3.3 Predicting the Binding Site Models of Novel Proteins

Given a novel Cys_2His_2 Zinc Finger protein, we first need to analyze its sequence and predict a binding site model. We first identify the four key residues at each DNA-binding domain, and then we utilize the learned sets of DNA-recognition preferences by assigning the appropriate probabilities and constructing a probabilistic model of the binding site. This is illustrated in Figure 5.

For example, Figure 6a compares the known binding site model of Sp1, to the one predicted using our approach. To prevent bias by known Sp1 sites in our training data, we apply a "*leave-protein-out*" cross-validation analysis, and predict the DNA-binding model of Sp1 using DNA-recognition preferences that were learned from a reduced dataset without Sp1 binding sequences.

3.4 *In Silico* Binding Experiments

We now use the predicted binding site models to scan genomic sequences for putative binding sites. Using the CIS algorithm [2], we score each possible binding position on the two DNA strands using a log-odds score (the log of the ratio between the probability of the binding site given the predicted model, and its probability given a 3^{rd}-order Markov model trained on genomic sequences). We then estimate the p-value of these scores and apply a *Bonferroni* correction to account for multiple tests within the same promoter region. Sites with a significant p-value (≤ 0.05 after Bonferroni correction) were marked as putative binding sites. For example, Figure 6b demonstrates such an *in silico* binding experiment for the human dihydrofolate reductase (DHFR) promoter, using the predicted binding site model of Sp1.

3.5 Quantitative Validation of Binding Site Predictions for Novel Proteins

To further evaluate the sets of recognition preferences, we mined the literature for experimentally verified binding sites of canonical Cys_2His_2 Zinc finger proteins. These include 43 binding sites, from 21 pairs of transcription factors and the natural genomic promoter regions of their target genes (some proteins have multiple binding sites per promoter). As described above, we utilized the learned DNA-recognition preferences to predict binding site models for the involved transcription factors, and used them to scan the respective promoter regions for putative binding sites. To ensure the validity of the test, we applied a "leave-protein-out" cross-validation test as specified above. Figure 6c summarizes these 21 *in silico* binding experiments using an ROC curve. Using $p = 0.05$ (with Bonferroni correction), our method marked 30 locations as putative binding sites, out of which 21 match experimental knowledge (sensitivity of 49% and specificity of 70%, hyper-geometric p-value $< 10^{-48}$).

3.6 Comparison with Other Computational Approaches

In a similar manner, we generated probabilistic binding site models for these transcription factors using the recognition preferences of Mandel-Gutfreund et al. [17] (that were used as a starting point in our learning algorithm), and repeated the quantitative analysis. As we show in Figure 6c, predictions based on these preferences have inferior accuracy.

In a recent study, Benos et al. [4] used *in vitro* specialized experimental data (such as SELEX and phage display) to assign position-specific DNA-recognition preferences for the $\mathrm{Cys_2His_2}$ Zinc Finger family (see detailed comparison of the two approaches in the Discussion section). As before, we used their preferences to generate probabilistic binding site models for these transcription factors, and then used them to scan the corresponding promoter regions. Once again, Figure 6c shows that predictions based on our sets of DNA-recognition preferences are more accurate.

3.7 Predictions Based on Genomic Data

To further evaluate our predictions on long genomics sequences, we used the binding locations of Sp1 along human Chromosomes 21 and 22, as mapped by an unbiased genome-wide chromatin immunoprecipitation (*ChIP*) assay [8]. We compiled two datasets of 1Kb-long sequences: one dataset included sequences that exhibited highly significant binding in the *ChIP* assay, while the other dataset included sequences that showed no binding at all (to be used as a control). We then performed *in silico* binding experiments using CIS [2], searching the sequences by the predicted binding site model of Sp1. Figure 6d compares the abundance of putative hits in both datasets. As can be seen, using a Bonferroni corrected threshold of 0.05, putative Sp1 binding sites were found in 45% of the experimentally-bound sequences, while only in 5% of the control sequences.

4 Discussion

In this paper we propose a general framework for predicting the DNA binding site models of novel transcription factors from known families. Our framework combines structural information about a DNA-binding family, with sequence data about binding sites for other proteins in the same family. We apply our approach to the canonical $\mathrm{Cys_2His_2}$ Zinc Finger DNA-binding family, and use a statistical estimation algorithm to derive a set of amino acid-nucleotide recognition preferences for each key position in the Zinc Finger DNA-binding domain. These recognition preferences can then be used to predict the binding site models of novel proteins from the same family. Finally, we use the predicted models to scan regulatory genomic regions of target genes, and identify their putative binding sites.

5 Prediction of Binding Sites Using Structure-Based Approaches

Structure-based approaches for prediction of transcription factor binding sites have recently gained much interest [4, 14, 17, 23, 25]. Most of the structural approaches define a protein-DNA binding model based on solved protein-DNA complexes, and attempt to identify DNA subsequences that fit best the amino acids that are determined as interacting with the DNA. While some of these studies [14, 19] used ensembles of solved protein-DNA complexes from all DNA-binding domains to extract general preferences for amino acid-base recognition, we and others focus on a single DNA-binding domain. Although less general, we hope that such an approach will lead to more fine-grained definitions of the binding preferences.

In a recent study, Benos et al. [4] assigned position-specific DNA-recognition potentials for the $\mathrm{Cys_2His_2}$ Zinc Finger family. Although the model they used is quite similar to ours, there are significant differences between the two. First, they relied only on aligned binding sites from *in vitro* specialized experiments, such as SELEX and phage display, to train their recognition preferences. Second, their assays screened artificial sequences of both artificial proteins and artificial DNA targets. In contrast, we rely on longer, unaligned natural binding data. Previous studies showed that there are discrepancies between SELEX-derived motifs and those derived from natural binding sites [21, 22]. As we showed, our sets of estimated DNA-recognition preferences are more consistent with independent experimental results [6, 9, 10, 28] and are superior to similar preferences derived by the other computational methods [4, 17]. To further illustrate this point, we returned the artificial binding sequences from TRANSFAC back into our training data, and obtained inferior predictions. Figure 6c summarizes a quantitative comparison between all models in identifying binding sites of novel proteins within genomic sequences. It should be stressed out that in order to prevent unfair bias, we use a "leave-protein-out" cross validation, hence removing all binding sites of a protein from the training data before testing it.

5.1 Analysis of the Estimated DNA-Recognition Preferences

A close examination of the learned sets of DNA-recognition preferences suggests that the protein-DNA recognition code is not deterministic, but rather spans a range of preferences. Moreover, our analyses show that a residue may have different nucleotide preferences depending on its context. For some amino acids, the qualitative preferences remain the same across various positions, while the quantitative preferences vary (e.g. arginine, see Figure 4). The DNA-binding preferences of other residues change across various positions. For example, histidine at position 3 tends to interact with guanine, while it shows no preference to any nucleotide at all other positions. Another example is the tendency of alanine at position 6 to face guanine. This preference, which was revealed automatically by our analysis, does not comply with both the chemical nature of alanine's side chain, nor with general examinations of amino acid-nucleotide interactions [14, 17]. We suspect that it is affected by the large fraction of Sp1 targets in our dataset. This potential interaction was implied before in Sp1 binding sites [5] and may reflect an interaction between the residue at position 2 with the complementary cytosine.

5.2 Inter-position Dependencies in the Binding Site

The Cys_2His_2 binding model inherently assumes that all positions within the binding site are independent of each other. This assumption is used in most computational approaches that model binding sites. Two papers [3, 7] discuss this issue in the context of the Cys_2His_2 Zinc Finger domain. Their analyses of binding affinity measurements suggest that some weak dependencies do exist among some positions of the binding sites of Egr-1. Nonetheless, a reasonable approximation of the binding specificities is obtained even when ignoring these dependencies. In another recent study [1], we evaluated probabilistic models that are capable of capturing such inter-position dependencies within binding sites. Our results showed that dependencies can be found in the binding sites of many proteins from various DNA-binding domains (especially from the helix-turn-helix and the homeo domains). However, our results also implied that using such models of dependencies in modeling the binding sites of Zinc Finger proteins does not lead to significant improvements [1]. Thus, we believe that the Cys_2His_2 binding model we use here is indeed a reasonable approximation of the actual binding.

5.3 Genome-wide Predictions of Binding Sites and Target Genes

In the current era, there is a growing gap between the number of known protein sequences and the number of experimentally verified binding sites. To better understand regulatory mechanisms in newly solved genomes, it is crucial to identify the direct target genes of novel DNA-binding proteins. Our method opens the way for such genome-wide assays. By predicting the binding site models of regulatory proteins, one might attempt to also classify the genes to those that contain significant binding sites at their regulatory promoter regions (hence, putative target genes) and those that do not. As we showed, our approach can scale up to such genome-wide scans.

5.4 Applications to Other DNA-Binding Domains

Theoretically, our approach can be extended to handle other structural families. In Figure 7, we analyze the number of binding sites needed for estimating the DNA-recognition preferences. We show that ~200 sites are sufficient for achieving similar likelihood values. Other possible families of DNA-binding domains, such as the leucine zipper, the homeodomain and the helix-turn-helix domain, have enough sites in TRANSFAC to allow similar analyses (1191, 505 and 201 sites, respectively). Unfortunately, this move requires that the various proteins

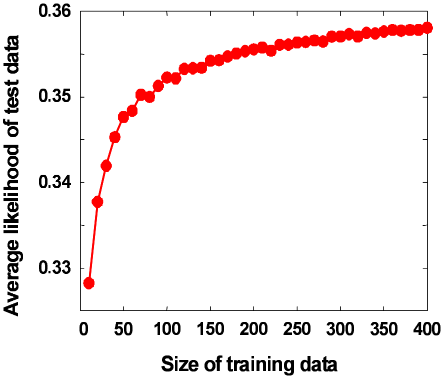


Fig. 7. Likelihood of held-out test data given different sizes of training datasets. The original data (455 canonical Cys_2His_2 zinc finger sites from TRANSFAC 7.3) were split into 10 equally-sized sets. We used each set as held-out test data, while applying the following procedure 10 times: various portions of different sizes (from 10 to 400 binding sites) were sampled from the remaining 90% of the data, and were used as training data for the EM algorithm. We then calculated the average likelihood of the held-out 10%

of the target DNA-binding domain will follow a common simple DNA-binding model. This is not the case for some families, where the binding models are far more flexible and complex. To handle these cases, more advanced models and learning techniques will be needed. Furthermore, for some families there is no simple way of inferring the interacting residues, based on the sequence of the protein (unlike the conserved Cys_2His_2 pattern in the Zinc Finger domain), and so the possible search space grows even further.

In spite of these drawbacks, we believe that structural approaches, as the one we show here, will lead to successful predictions of binding site models, and following that, to accurate identification of the target genes of novel proteins, even on genome-wide scales. Eventually, such approaches will be utilized to reconstruct larger and larger portions of the transcriptional regulatory networks that control the living cell.

Acknowledgments

The authors wish to thank Yael Altuvia, Ernest Fraenkel, Benjamin Gordon, Ruth Hershberg, Dalit May, Lena Nekludova, Aviv Regev, and Eran Segal for helpful discussions. T.K. is supported by the Horowitz Fellowship. This work was supported in part by grants from the Israeli Ministry of Science and the Israeli Science Foundation (administered by the Israeli Academy of Sciences and Humanities).

Availability

A web-server for predicting the binding sites of Cys_2His_2 Zinc Finger proteins, based on their sequences and on the estimated recognition preferences, can be accessed at `http://compbio.cs.huji.ac.il/predictBS`.

References

1. Barash, Y., *et al.*: Modeling dependencies in Protein-DNA binding sites. Proc. of the 7th International Conf. on Research in Computational Molecular Biology (2003) 28–37
2. Barash, Y., *et al.*: CIS: Compound Importance Sampling method for protein-DNA binding site p-value estimation. Bioinformatics (2004)
3. Benos, P.V., Bulyk, M.L., Stormo, G.D.: Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. **30** (2002) 4442–4451
4. Benos, P.V., Lapedes, A.S., Stormo, G.D.: Probabilistic code for DNA recognition by proteins of the EGR family. J. Mol. Biol. **323** (2002) 701–727
5. Berg, J.M.: Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. Proc. Natl. Acad. Sci. USA **89** (1992) 11109–11110
6. Bulyk, M.L., *et al.*: Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc. Natl. Acad. Sci. USA **98** (2001) 7158–7163
7. Bulyk, M.L., Johnson, P.L.F., Church, G.M.: Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. **30** (2002) 1255–1261
8. Cawley, S., *et al.*: Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell **116**(4) (2004) 499–509
9. Choo, Y., Klug, A.: Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. Proc. Natl. Acad. Sci. USA **91** (1994) 11168–11172
10. Choo, Y., Klug, A.: Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc. Natl. Acad. Sci. USA **91** (1994) 11163–11167
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood form incomplete data via the EM algorithm. J. Royal Stat. Soc. B. **39** (1977) 1–38
12. Eddy, S.R.: Profile hidden Markov models. Bioinformatics **14** (1998) 755–763
13. Elrod-Erickson, M., Benson, T.E., Pabo, C.O.: High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. Structure **6** (1998) 451–464
14. Kono, H., Sarai, A.: Structure-based prediction of DNA target sites by regulatory proteins. Proteins **35** (1999) 114–131
15. Kriwacki, R.W., *et al.*: Sequence-specific recognition of DNA by zinc-finger peptides derived from the transcription factor Sp1. Proc. Natl. Acad. Sci. USA **89** (1992) 9759–9763
16. Luscombe, N.M., Laskowski, R.A., Thornton, J.M.: Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. **29** (2001) 2860–2874
17. Mandel-Gutfreund, Y., Baron, A., Margalit, H.: A structure-based approach for prediction of protein binding sites in gene upstream regions. Proc. of the Pac. Symp. Biocomput. (2001) 139–150
18. Mandel-Gutfreund, Y., Schueler, O., Margalit, H.: Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J Mol Biol. 253 (1995) 370–382
19. Mandel-Gutfreund, Y., Margalit, H.: Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. Nucleic Acids Res. **26** (1998) 2306–2312

20. Pavletich, N.P., Pabo, C.O.: Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. Science **252** (1991) 809–817
21. Robison, K., McGuire, A.M., Church, G.M.: A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. J. Mol. Biol. **284** (1998) 241–254
22. Shultzaberger, R.K., Schneider, T.D.: Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. Nucleic Acids Res. **27** (1999) 882–887
23. Steffen, N.R., *et al.*: DNA sequence and structure: direct and indirect recognition in protein-DNA binding. Bioinformatics **18 Suppl 1** (2002) S22–S30
24. Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics **16**(1) (2000) 16–23
25. Suzuki, M., Gerstein, M., Yagi, N.: Stereochemical basis of DNA recognition by Zn fingers. Nucleic Acids Res. **22** (1994) 3397–3405
26. Tupler, R., Perini, G., Green, M.R.: Expressing the human genome. Nature **409**(6822) (2001) 832–833
27. Wingender, E., *et al.*: The TRANSFAC system on gene expression regulation. Nucleic Acids Res. **29** (2001) 281–283
28. Wolfe, S.A., *et al.*: Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. J. Mol. Biol. **285** (1999) 1917–1934